Application to Microarrays

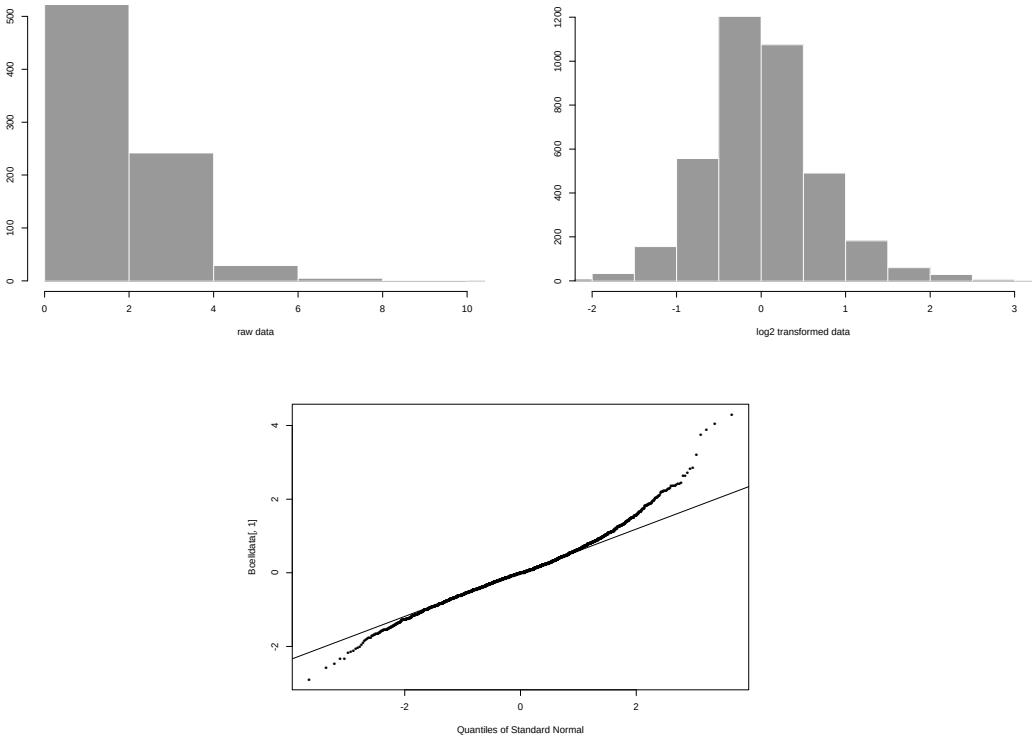
Application to small case/control data sets.

With: Ying Qi, SAIC/NCI-Fredrick

The semiparametric method is applied to $\log_2$ microarray gene-expression data from two case and control groups. We use $h(x) = x$ since the $\log_2$ appear symmetric.

The method holds the distribution of one group as "reference", and combines the data from both groups in the estimation of the reference distribution and its "distortion", and in testing the hypothesis of distribution equality. The method is useful for gene classification and identification.

Histograms of the combined microarray GCAC4026 data. Top left, raw data. Top right, $\log_2$-transformed data. Bottom, (qq-plot) comparison of the quantiles of the $\log_2$ data with the quantiles of the standard normal distribution.

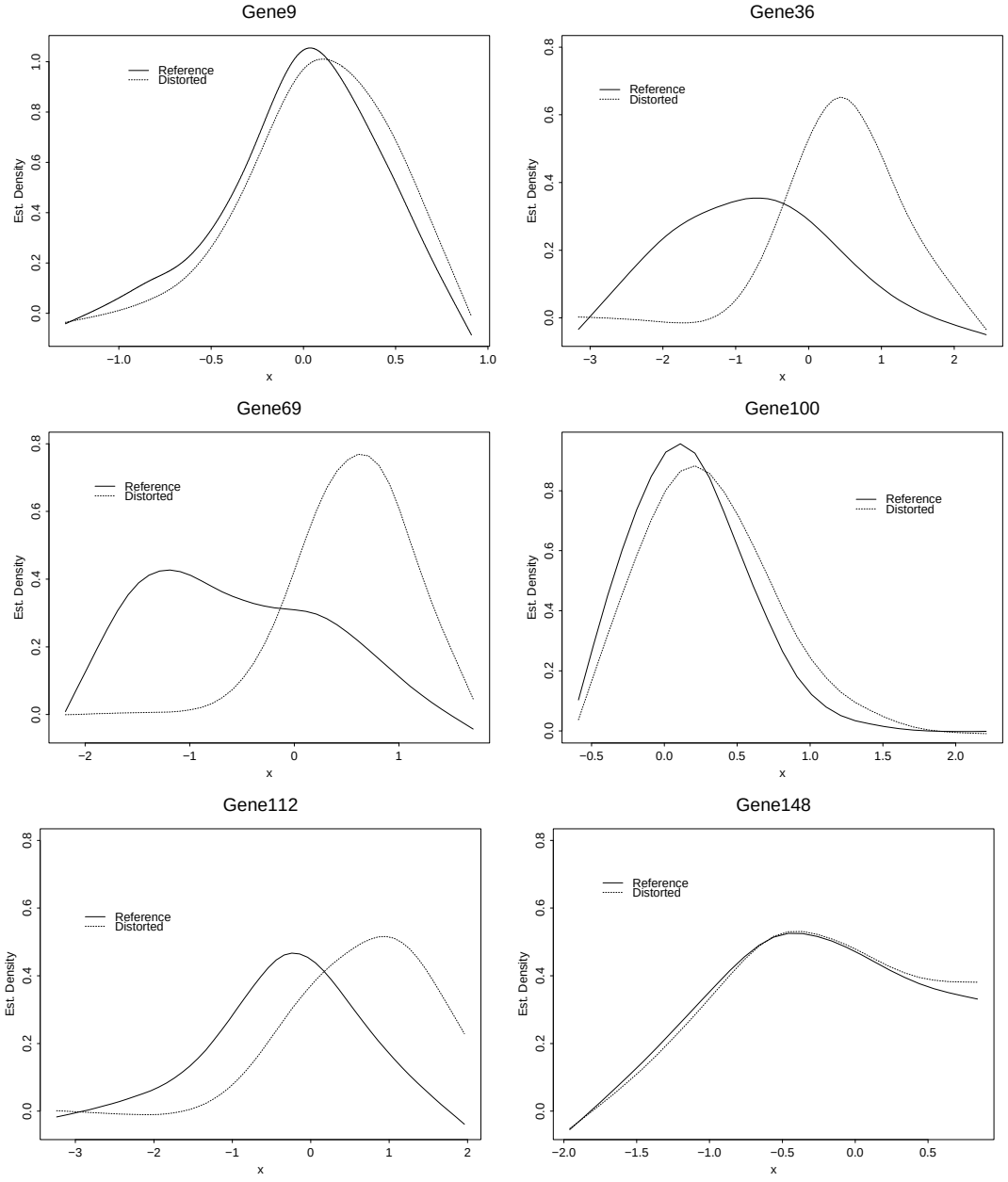


Summary of the semiparametric analysis for the indicated genes from GC148 and AC148 (reference).
$m = 2, q = 1$. $21 \leq n_1, n_2 \leq 24$.

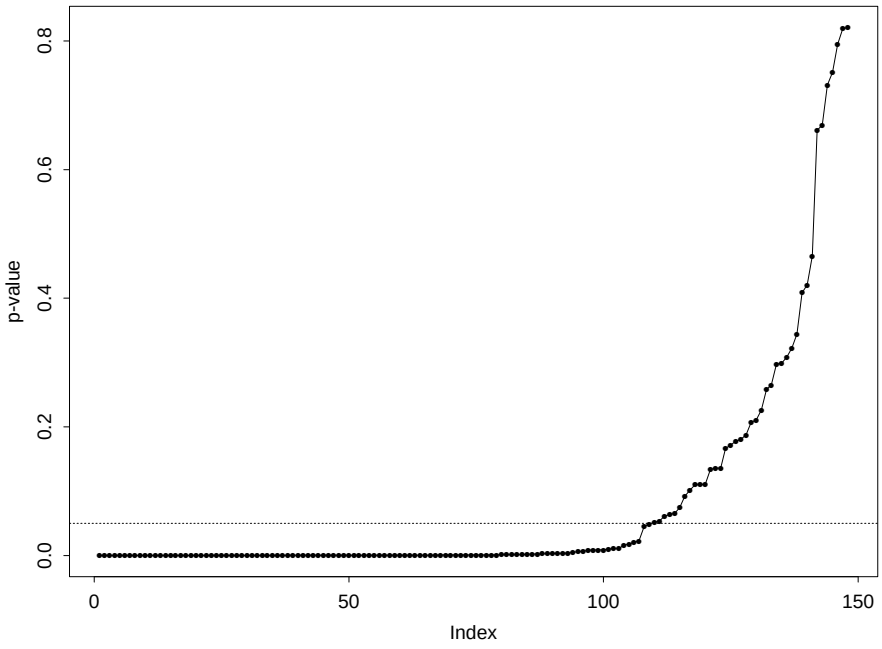
Gene	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\mathcal{X}_1$	p-value
9	-0.049	0.810	1.464	0.226
36	0.452	3.199	97.170	0.000
69	-0.073	2.415	47.353	0.000
100	-0.159	0.751	1.244	0.265
112	-0.241	1.698	31.857	0.000
148	0.030	0.100	0.051	0.821

Gene	$\widehat{\alpha}_1$	$SE(\widehat{\alpha}_1)$	$\widehat{\beta}_1$	$SE(\widehat{\beta}_1)$
9	-0.049	0.068	0.810	0.740
36	0.452	0.353	3.199	1.028
69	-0.073	0.301	2.415	0.712
100	-0.159	0.137	0.751	0.637
112	-0.241	0.212	1.698	0.537
148	0.030	0.134	0.100	0.444

Smoothed estimated reference $g(x)$ and its distortion for indicated genes.



Sorted 148 p-values obtained from the $\mathcal{X}_1$ test comparing the gene behavior in the AG and AC groups. The equidistribution hypothesis is rejected for all 109 p-values below the dashed line marking the 0.05 level.



Scatterplot and a smooth approximation of $\sqrt{\mathcal{X}_1}$ versus the corresponding $|t|$ values.